Towards Spatio-Temporal Face Alignment in Unconstrained Conditions

Romain Belmonte, Nacim Ihaddadene, Pierre Tirilly, Chaabane Djeraba, Ioan Marius Bilasco

Towards Spatio-Temporal Face Alignment in Unconstrained Conditions

R. Belmonte[1,3], N. Ihaddadene[1], P. Tirilly[2], M. Bilasco[3] and C. Djeraba[2]
[1]*ISEN Lille, Yncrea Hauts-de-France, France*
[2]*Univ. Lille, CNRS, Centrale Lille, IMT Lille Douai, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France*
[3]*Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France*
{*romain.belmonte, nacim.ihaddadene*}*@yncrea.fr*, {*pierre.tirilly, marius.bilasco, chaabane.djeraba*}*@univ-lille1.fr*

Keywords: Face Alignment, Spatio-Temporal Models, Landmark Localization, Video-based, Unconstrained Conditions

Abstract: Face alignment is an essential task for many applications. Its objective is to locate feature points on the face, in order to identify its geometric structure. Under unconstrained conditions, the different variations that may occur in the visual context, together with the instability of face detection, make it a difficult problem to solve. While many methods have been proposed, their performances under these constraints are still not satisfactory. In this article, we claim that face alignment should be studied using image sequences rather than still images, as it has been done so far. We show the importance of taking into consideration the temporal information under unconstrained conditions.

1 INTRODUCTION

The problem of face alignment, also called facial landmark localization, has raised much interest and experienced rapid progress in recent years (Jin and Tan, 2016). Given the position and size of a face, the alignment process, illustrated in Figure 1, consists in determining the geometry of the face components containing the most useful information (e.g., eyes, nose, mouth). This ability to model non-rigid facial structures is now used in various fields such as face analysis (e.g., identification, expression recognition) (Sun et al., 2014), human-computer interaction (Akakin and Sankur, 2009) or information retrieval (Park and Jain, 2010). However, despite the large number of methods available in the literature, the performance of face alignment under unconstrained conditions remains limited (Sagonas et al., 2016).

Even today, this problem continues to be studied using still images (Jin and Tan, 2016). Yet, due to the ubiquity of video sensors, the vast majority of applications rely on image sequences. In addition, many tasks related to face analysis or, more broadly, human behavior analysis have leveraged temporal information (Simonyan and Zisserman, 2014; Fan et al., 2016). The first survey on face alignment have already suggested to study the problem using image sequences, but without providing actual arguments (Çeliktutan et al., 2013). Our motivation in this paper is to show that taking temporal information into account for this problem could greatly contribute to improve performance under unconstrained conditions.

This article is structured as follows: in Section 2, we describe the reasons that led us towards spatio-temporal approaches for face alignment. In particular, we review the existing solutions and put into perspective the performance of the most recent methods. Section 3 shows the benefits of temporal information under unconstrained conditions. Finally, we conclude with Section 4.

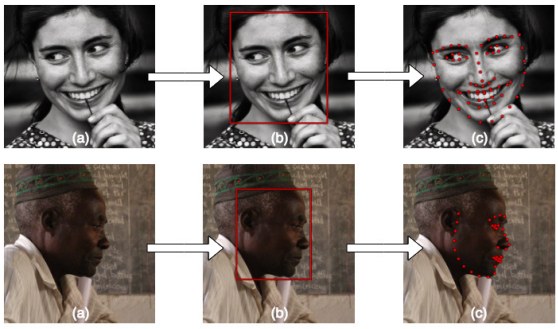

Figure 1: Face alignment process: (a) original image, (b) face detection, (c) face alignment. Images from Menpo (Zafeiriou et al., 2017).

Type	Datasets	#Images	#Landmarks
Static	AFLW (Köstinger et al., 2011)	25,993	21
	AFW (Zhu and Ramanan, 2012)	250	6
	HELEN (Le et al., 2012)	2,330	194
	iBug (Sagonas et al., 2013)	135	68
	LFPW (Belhumeur et al., 2013)	1,432	35
	COFW (Burgos-Artizzu et al., 2013)	1,007	29
	300W (Sagonas et al., 2016)	3,837	68
	300W-LP (Zhu et al., 2016)	61,225	68
	MENPO (Zafeiriou et al., 2017)	7564	68/39
Dynamic	300VW (Shen et al., 2015)	218,595	68

Table 1: Datasets captured under unconstrained conditions.

2 LITERATURE REVIEW

In this section, the datasets and evaluation metrics used for face alignment under unconstrained conditions are discussed. The main categories of methods from literature are reviewed and their performance is analyzed.

2.1 Datasets and Metrics

In recent years, many datasets for face alignment have been made available to the scientific community (cf. Table 1). The images included in these datasets are collected on social networks or image search services such as Google, Flickr, or Facebook, bringing more realism to the data. The annotation is performed either manually or semi-automatically, sometimes with the help of the Amazon Mechanical Turk platform. The quality of the annotations, however, may vary (Sagonas et al., 2016; Bulat and Tzimiropoulos, 2017; Çeliktutan et al., 2013). The annotation scheme used (i.e., position and number of landmarks) may also differ from one dataset to another. Currently, the scheme composed of 68 landmarks (Gross et al., 2010), illustrated in Figure 1, is the most widely used. Since this scheme does not fit at extreme poses, a 39-point-based scheme has recently been proposed for profile faces (Zafeiriou et al., 2017) (see Figure 1).

Today, due in particular to the emergence of deep learning techniques, it may be necessary to have a large number of annotated images, which existing datasets do not necessarily provide. In the literature, various augmentation methods are used to circumvent this problem. Some operations (e.g., rotation, mirroring, disturbance of the detection window position and size) can be applied to images to generate new training samples. Other more complex processes such as the generation of synthetic images may also be used (Zhu et al., 2016).

Moreover, it is crucial to have data that is representative of the problem in order to answer it. Static datasets do not cover all the difficulties encountered by applications. The constraints related to the movements of persons or cameras are not currently considered. A dataset composed of image sequences captured under unconstrained conditions has recently been published by (Shen et al., 2015) (cf. Figure 2). This data, in addition to being more realistic, provides some clues in favor of the use of temporal information for face alignment.

Figure 2: Challenges encountered under unconstrained conditions: occlusions (b), (d), (f), pose (a), (e), illumination (a), (b), expressions (c). Images from 300VW (Shen et al., 2015).

To evaluate predictions on this data, the mean square error normalized by the interocular distance (NMSE) is generally used. Beyond an error value of 8%, landmarks are mostly not located correctly and the prediction is considered as a failure. Normalization by the interocular distance, although not very robust to extreme poses, is the most widespread. Other normalization factors are sometimes used, such as the diagonal of the detection window. On a set of images, average NMSE is the simplest and most intuitive evaluation metric. However, it can be strongly affected by a few outliers. A graphical representation of the error distribution function is therefore increasingly used. It corresponds to the proportion of images for which the error is less than or equal to a certain threshold (e.g.,

Method	300W-A	300W-B	300W
RCPR (Burgos-Artizzu et al., 2013)	6.18	17.26	8.35
SDM (Xiong and De la Torre, 2013)	5.57	15.40	7.50
ESR (Cao et al., 2014)	5.28	17.00	7.58
CFAN (Zhang et al., 2014)	5.50	16.78	7.69
LBF (Ren et al., 2014)	4.95	11.98	6.32
CFSS (Zhu et al., 2015)	4.73	9.98	5.76
RPPE (Yang et al., 2015a)	5.50	11.57	6.69
3DDFA (Zhu et al., 2016)	6.15	10.59	7.01
TCDCN (Zhang et al., 2016)	4.80	8.60	5.54
RAR (Xiao et al., 2016)	4.12	**8.35**	**4.94**
RCFA (Wang et al., 2016)	**4.03**	9.85	5.32
R-DSSD (Liu et al., 2017a)	4.16	9.20	5.59

Table 2: Performances of recent literature methods. Normalized mean squared error is reported.

8%). The area under the curve and the failure rate, that is, the percentage of images for which the error is greater than the threshold, are sometimes calculated from this representation.

2.2 Registration Solutions

In the literature, two main categories of methods to locate facial landmarks are proposed. First, there are generative methods which rely on joint parametric models of appearance and shape (Cootes et al., 2001). The alignment is formulated as an optimization problem with the objective of finding the parameters allowing the best possible instance of the model for a given face. The appearance can be represented holistically or locally, using regions of interest centered on the landmarks.

Then, there are discriminative methods which infer the position of the landmarks directly from the appearance of the face. They either learn independent local detectors or regressors for each landmark associated with a shape model that regularizes the predictions (Saragih et al., 2011), or one or more vector regression functions able to infer all the landmarks and implicitly include a shape constraint (Xiong and De la Torre, 2013; Cao et al., 2014; Ren et al., 2014; Zhu et al., 2015). In this category, methods based on deep learning (e.g., convolutional neural networks, autoencoders) have recently led to a significant improvement in performance under unconstrained conditions, notably through their ability to model non-linearity and learn problem-specific features (Xiao et al., 2016; Wang et al., 2016; Liu et al., 2017a).

While most methods address the problem globally, some focus specifically on a single challenge (Burgos-Artizzu et al., 2013; Yang et al., 2015a; Zhu et al., 2016). (Burgos-Artizzu et al., 2013) explicitly model the occlusions and show that this additional information helps to improve the estimation of landmarks positions under unconstrained conditions. Training, however, requires to annotate occlusions. (Zhu et al., 2016) focus on extreme poses and propose to infer a 3D dense model rather than a sparse 2D model. Their method is able to handle horizontal variations ranging from -90°to 90°.

Others suggest that face alignment should not be treated as an independent problem and propose to jointly learn various related tasks in order to achieve individual performance gains (Ranjan et al., 2016; Zhang et al., 2016). In the work of (Zhang et al., 2016), alignment is learned in conjunction with pose estimation, gender recognition, facial expressions recognition, and the appearance of facial attributes. However, this type of approach can make the training stage much more complex because the convergence rates may vary from one task to another.

The performance of current face alignment methods are referenced in Table 2. These methods were evaluated on the 300W dataset, composed of categories of variable difficulty. The 300-A category corresponds to images that do not include strong constraints. Category 300-B contains more complex images with large variations in pose and expression, as well as occlusions. We note that for category B the average error is more than twice the one obtained on category A.

Despite the quantity of methods proposed in the literature and the recent major advances, we can see from these results that the problems encountered under unconstrained conditions are still far from being solved. Because of their significant influence on facial appearance, variations in pose and occlusions are among the most difficult challenges (Shen et al., 2015; Sagonas et al., 2016). We show in Section 3 how temporal approaches could help to address these problems.

Method	Category 1		Category 2		Category 3	
	AUC	FR(%)	AUC	FR(%)	AUC	FR(%)
(Uricár et al., 2015)	0.657	7.622	0.677	4.131	0.574	7.957
(Xiao et al., 2015)	0.760	5.899	0.782	3.845	0.695	7.379
(Rajamanoharan and Cootes, 2015)	0.735	6.557	0.717	3.906	0.659	8.289
(Wu and Ji, 2015)	0.674	13.925	0.732	5.601	0.602	13.161
(Zhu et al., 2015; Danelljan et al., 2015)	0.729	6.849	0.777	0.167	0.684	8.242
(Yang et al., 2015c)	**0.791**	**2.400**	**0.788**	**0.322**	**0.710**	**4.461**

Table 3: Comparison of methods from (Shen et al., 2015; Chrysos et al., 2017), on the 3 categories of 300VW. Area under the curve (AUC) and failure rate (FR) are reported.

3 THE BENEFITS OF TEMPORAL INFORMATION

In this section, we show how temporal information can be beneficial to the problem of face alignment under unconstrained conditions. Issues raised by the dependence to face detection are first discussed. Constraints including the trajectories of the landmarks are then pointed out as solutions to enhance the robustness and quality of face alignment.

3.1 Rigid and Non-rigid Tracking

Face detection under unconstrained conditions is a complex problem to solve (Zafeiriou et al., 2015). Given its role in face alignment, (Yang et al., 2015b) studied the dependence between these two tasks. They showed a high sensitivity of alignment to detection quality. Thus, besides detection failures, factors such as variations in the scale and position of the detection window may disrupt the alignment.

A solution to avoid the dependence on face detection is to perform non-rigid face tracking. (Shen et al., 2015) recently proposed a comparative analysis of current non-rigid face tracking methods. The results are referenced in Table 3. The most popular strategy is tracking by detection, that is, face detection and alignment on each image independently, without making use of adjacent frames. An alternative to tracking by detection is to use a substitute for the detection such as a generic (i.e., rigid) tracking algorithm (Danelljan et al., 2015). One of the advantages of generic tracking algorithms is that they are able to take into account some variations in the appearance of the target object during tracking (Kristan et al., 2016). (Chrysos et al., 2017) evaluate this strategy and compare it to tracking by detection. In general, generic tracking makes it possible to be more robust to the challenges encountered under unconstrained conditions. It is, however, probable that, as with detection, alignment is sensitive to changes in the tracking window.

(Yang et al., 2015c) avoid face detection at each frame or rigid tracking by proposing a spatio-temporal cascade regression. They initialize the shape at the current frame from the similarity parameters at the previous frame. They incorporate a re-initialization mechanism based on the quality of the prediction in order to avoid any drift in the alignment. Their method can greatly reduce the failure rate while improving overall performance (cf. Table 3). (Sánchez-Lozano et al., 2016) propose an incremental cascaded continuous regression. In contrast to (Yang et al., 2015c) which retains a generic model after learning, here a pre-trained model is updated online to become specific to each person during tracking. This type of approach yields better results than a generic model. In the end, non-rigid tracking produces more accurate fitting than tracking by detection. It takes advantage of adjacent frames to improve initialization and variations in appearance to increasingly become person-specific. Yet, other information such as the trajectories of the landmarks through the image sequence seems relevant to consider (Hamarneh and Gustavsson, 2001). This will be discussed in detail in the next subsection.

3.2 Additional Constraints

Whether explicit or implicit, the shape constraints present in most alignment methods are crucial to obtain good performance in unconstrained conditions. In image sequences, an additional constraint may be applied to the trajectories of the landmarks. Bayesian filters such as Kalman filters or particle filters can be used for this purpose. However, (De and Kautz, 2017) highlighted the marginal gain provided by these approaches and showed the superiority of recurrent neural networks for dynamic face analysis.

(Peng et al., 2016; Hou et al., 2017) also use recurrent neural networks to exploit the dynamic features of the face. They compare a recurrent and a non-recurrent version of their network and show that recurrent learning improves the stability of predictions

and the robustness to occlusions, variations in pose and expressions. The performances of their methods are referenced in Table 4. Taking into account temporal information results in a decrease of the average error of more than 1% compared to state-of-the art static methods.

Approach	Method	300VW
Static	(Zhang et al., 2016)	7.59
Dynamic	(Peng et al., 2016)	6.25
	(De and Kautz, 2017)	6.16
	(Liu et al., 2017b)	**5.59**

Table 4: Comparison of a static multi-task method (Zhang et al., 2016), with three dynamic methods based on recurrent neural networks (Peng et al., 2016; De and Kautz, 2017; Liu et al., 2017b). Normalized mean squared error is reported.

More recently, (Liu et al., 2017b) propose a two-stream recurrent network composed of a spatial stream that preserve the holistic facial shape structure and a temporal stream that discover shape-sensitive and spatio-temporal features. Their method outperforms those based on a single stream (see Table 4). These are only the beginnings of the use of temporal information for the problem of face alignment. Recurrent neural networks, although advantageous, are capable of characterizing only the global motion. In other related tasks, local motion (i.e., over a few frames) sometimes associated with global motion has led to interesting results (Hasani and Mahoor, 2017; Fan et al., 2016) and could be equally beneficial to face alignment.

4 CONCLUSION

In this paper we presented a review of current work on face alignment under unconstrained conditions. This problem has been studied on still images for several decades, despite applications being mainly based on image sequences. Moreover, many tasks related to face analysis or, more broadly, human behavior analysis have leveraged temporal information. To the best of our knowledge, this is one of the first surveys to quantify the difference between static and dynamic face alignment methods. Especially, we have shown that taking into consideration the temporal information greatly contribute to overcome unconstrained conditions challenges. Recent work that exploit the dynamic features of the face are able to improve initialization and stability of predictions leading to more accurate fitting. Nevertheless, there is still much to be done.

REFERENCES

Akakin, H. C. and Sankur, B. (2009). Analysis of head and facial gestures using facial landmark trajectories. In *European Workshop on Biometrics and Identity Management*, pages 105–113.

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., and Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 35(12):2930–2940.

Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). *arXiv preprint arXiv:1703.07332*.

Burgos-Artizzu, X. P., Perona, P., and Dollár, P. (2013). Robust face landmark estimation under occlusion. In *ICCV*, pages 1513–1520.

Cao, X., Wei, Y., Wen, F., and Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190.

Çeliktutan, O., Ulukaya, S., and Sankur, B. (2013). A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, 2013(1):13.

Chrysos, G. G., Antonakos, E., Snape, P., Asthana, A., and Zafeiriou, S. (2017). A comprehensive performance evaluation of deformable face tracking "in-the-wild". *International Journal of Computer Vision*, pages 1–35.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685.

Danelljan, M., Hager, G., Shahbaz Khan, F., and Felsberg, M. (2015). Learning spatially regularized correlation filters for visual tracking. In *ICCV*, pages 4310–4318.

De, J. G. X. Y. S. and Kautz, M. J. (2017). Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *CVPR*.

Fan, Y., Lu, X., Li, D., and Liu, Y. (2016). Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *International Conference on Multimodal Interaction*, pages 445–450.

Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). Multi-PIE. *Image and Vision Computing*, 28(5):807–813.

Hamarneh, G. and Gustavsson, T. (2001). Deformable spatio-temporal shape models: Extending ASM to 2D+ time. In *BMVC*, pages 1–10.

Hasani, B. and Mahoor, M. H. (2017). Facial expression recognition using enhanced deep 3D convolutional neural networks. *arXiv preprint arXiv:1705.07871*.

Hou, Q., Wang, J., Bai, R., Zhou, S., and Gong, Y. (2017). Face alignment recurrent network. *Pattern Recognition*.

Jin, X. and Tan, X. (2016). Face alignment in-the-wild: A survey. *arXiv preprint arXiv:1608.04188*.

Köstinger, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2011). Annotated facial landmarks in the wild: A

large-scale, real-world database for facial landmark localization. In *ICCV Workshops*, pages 2144–2151.

Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Čehovin, L., Vojir, T., Häger, G., Lukežič, A., and Fernandez, G. (2016). The visual object tracking vot2016 challenge results. In *ECCV Workshops*, pages 777–823, Cham.

Le, V., Brandt, J., Lin, Z., Bourdev, L., and Huang, T. (2012). Interactive facial feature localization. *ECCV*, pages 679–692.

Liu, H., Lu, J., Feng, J., and Zhou, J. (2017a). Learning deep sharable and structural detectors for face alignment. *IEEE Transactions on Image Processing*, 26(4):1666–1678.

Liu, H., Lu, J., Feng, J., and Zhou, J. (2017b). Two-stream transformer networks for video-based face alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Park, U. and Jain, A. K. (2010). Face matching and retrieval using soft biometrics. *IEEE Transactions on Information Forensics and Security*, 5(3):406–415.

Peng, X., Feris, R. S., Wang, X., and Metaxas, D. N. (2016). A recurrent encoder-decoder network for sequential face alignment. In *ECCV*, pages 38–56.

Rajamanoharan, G. and Cootes, T. F. (2015). Multi-view constrained local models for large head angle facial tracking. In *ICCV Workshops*, pages 18–25.

Ranjan, R., Patel, V. M., and Chellappa, R. (2016). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *arXiv preprint arXiv:1603.01249*.

Ren, S., Cao, X., Wei, Y., and Sun, J. (2014). Face alignment at 3,000 fps via regressing local binary features. In *CVPR*, pages 1685–1692.

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2013). A semi-automatic methodology for facial landmark annotation. In *CVPR Workshops*, pages 896–903.

Sánchez-Lozano, E., Martinez, B., Tzimiropoulos, G., and Valstar, M. (2016). Cascaded continuous regression for real-time incremental face tracking. In *ECCV*, pages 645–661.

Saragih, J. M., Lucey, S., and Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215.

Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., and Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCV Workshops*, pages 1003–1011.

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576.

Sun, Y., Wang, X., and Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In *CVPR*, pages 1891–1898.

Uricár, M., Franc, V., and Hlavác, V. (2015). Facial landmark tracking by tree-based deformable part model based detector. In *ICCV Workshops*, pages 10–17.

Wang, W., Tulyakov, S., and Sebe, N. (2016). Recurrent convolutional face alignment. In *ACCV*, pages 104–120.

Wu, Y. and Ji, Q. (2015). Shape augmented regression method for face alignment. In *ICCV Workshops*, pages 26–32.

Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., and Kassim, A. (2016). Robust facial landmark detection via recurrent attentive-refinement networks. In *ECCV*, pages 57–72.

Xiao, S., Yan, S., and Kassim, A. A. (2015). Facial landmark detection via progressive initialization. In *ICCV Workshops*, pages 33–40.

Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539.

Yang, H., He, X., Jia, X., and Patras, I. (2015a). Robust face alignment under occlusion via regional predictive power estimation. *IEEE Transactions on Image Processing*, 24(8):2393–2403.

Yang, H., Jia, X., Loy, C. C., and Robinson, P. (2015b). An empirical study of recent face alignment methods. *arXiv preprint arXiv:1511.05049*.

Yang, J., Deng, J., Zhang, K., and Liu, Q. (2015c). Facial shape tracking via spatio-temporal cascade shape regression. In *ICCV Workshops*, pages 41–49.

Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., and Shen, J. (2017). The menpo facial landmark localisation challenge: A step towards the solution. In *CVPR Workshops*.

Zafeiriou, S., Zhang, C., and Zhang, Z. (2015). A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 138:1–24.

Zhang, J., Shan, S., Kan, M., and Chen, X. (2014). Coarse-to-fine auto-encoder networks for real-time face alignment. In *ECCV*, pages 1–16.

Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2016). Learning deep representation for face alignment with auxiliary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):918–930.

Zhu, S., Li, C., Change Loy, C., and Tang, X. (2015). Face alignment by coarse-to-fine shape searching. In *CVPR*, pages 4998–5006.

Zhu, X., Lei, Z., Liu, X., Shi, H., and Li, S. Z. (2016). Face alignment across large poses: A 3D solution. In *CVPR*, pages 146–155.

Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886.